second edition

Fundamentals of Biostatistics

Bernard Rosner

second edition

Fundamentals of Biostatistics

Bernard Rosner

Harvard University

Duxbury Press

Boston, Massachusetts

PWS PUBLISHERS

Prindle, Weber & Schmidt • Duxbury Press • PWS Engineering • Breton Publishers •
20 Park Plaza • Boston, Massachusetts 02116

PWS Publishers is a division of Wadsworth, Inc.

Library of Congress Cataloging-in-Publication Data
Rosner, Bernard (Bernard A.)
Fundamentals of biostatistics.

Includes bibliographies and index.
1. Biometry. 2. Medical statistics. I. Title.
QH323.5.R674 1986 574'.072 85-13088
ISBN 0-87150-981-4

ISBN 0-87150-981-4

Printed in the United States of America
86 87 88 89 90 — 10 9 8 7 6 5 4 3 2 1

Editor: Michael Payne
Production Coordinator and designer: S. London
Manuscript Editor: Ian List
Typesetting: Desmond Doyle Phototypesetters
Printing and Binding Cover and Text: R.R. Donnelley & Sons Company

Preface

I have written this introductory-level biostatistics text for upper-level undergraduate or graduate students interested in medicine or other health-related areas. This book requires no previous background in statistics, and its mathematical level assumes only a knowledge of algebra.

Fundamentals of Biostatistics evolved from a set of notes that I used in a course in biostatistics taught to Harvard University undergraduates and Harvard Medical School students over the past ten years. I wrote this book to help motivate students to master those statistical methods that are most often used in the medical literature. It is important from the student's viewpoint that the example material used to develop these methods be representative of what actually exists in the literature. Therefore most examples and exercises used in this book are either based on actual articles from the medical literature or on actual medical research problems that I have encountered during my consulting experience at the Harvard Medical School.

Most other introductory statistics texts either use a completely nonmathematical, cookbook approach or develop the material in a rigorous, sophisticated mathematical framework. In this book I have attempted to follow an intermediate course, minimizing the amount of mathematical formulation and yet giving complete explanations of all the important concepts. *Every* new concept is developed systematically through completely worked examples from current medical research problems. In addition, computer output is introduced where appropriate to illustrate these concepts.

The material in this book is suitable for either a one- or two-semester course in biostatistics. The material in Chapters 1 through 8 and Chapter 10 is suitable for a one-semester course. The instructor may select appropriate material from the other chapters as time permits.

The following changes have been made in the second edition:

- The number of exercises has more than doubled from the first edition, to over 1000 exercises overall. In particular, over 300 "drill" exercises have been added to facilitate immediate student comprehension of the material.
- The treatment of EDA has been augmented, including the presentation of box plots.

•Tables are supplied for the binomial and Poisson distributions to minimize computational complexity when working with these distributions.

•Extensive computer output is provided to reinforce the crucial concept of a sampling distribution. The treatment of the central limit theorem is handled in a similar fashion.

•A more thorough coverage is given to the important notions of estimation of power and sample size for one-sample problems in Chapter 7 and for two-sample problems in Chapters 8 and 10.

•The notion of an odds ratio for 2 × 2 tables is introduced, including appropriate methods of estimation and hypothesis testing.

•The Mantel-Haenszel test is introduced for the combination of data from more than one 2 × 2 table.

• The Kappa statistic is presented as a method for measuring reproducibility for discrete data.

•An introduction is given to multiple logistic regression as an analogue to multiple linear regression for binary outcome variables. Extensive computer output is used to motivate each of these methods.

•The Kruskal-Wallis test is presented as a nonparametric alternative to the one-way analysis of variance.

Fundamentals of Biostatistics, second edition, is organized as follows:

Chapter 1 is an *introductory chapter* giving an outline of the development of an actual medical study that I was involved with. It provides a unique sense of the role of biostatistics in the medical research process.

Chapter 2 concerns *descriptive statistics* and presents all the major numerical and graphical tools used for displaying medical data. This chapter is especially important for both consumers and producers of medical literature, since much of the actual communication of information is accomplished via descriptive material.

Chapters 3 through 5 discuss *probability*. The basic principles of probability are developed, and the most common probability distributions, such as the binomial and normal distributions, are introduced. These distributions are used extensively in the later chapters of the book.

Chapters 6 through 10 cover some of the basic methods of *statistical inference*.

Chapter 6 introduces the concept of drawing random samples from populations. The difficult notion of a sampling distribution is also developed, including an introduction to the most common sampling distributions, such as the t and chi-square distributions. The basic methods of *estimation* are also presented, including an extensive discussion of confidence intervals.

Chapters 7 and 8 contain the basic principles of *hypothesis testing*. The most elementary hypothesis tests for normally distributed data, such as the t test, are also fully discussed for the one- and two-sample problems.

Chapter 9 covers the basic principles of *nonparametric statistics*. The assumptions of normality are relaxed, and distribution-free analogues are developed for the tests in Chapters 7 and 8.

Chapter 10 contains the basic concepts of *hypothesis testing* as applied to categorical data, including some of the most widely used statistical procedures, such as the chi-square test and Fisher's exact test.

Chapter 11 develops the principles of *regression analysis*. The case of simple linear regression is thoroughly covered, and extensions are provided for the multiple regression case. An important section on the limitations of the use of regression analysis is also included.

Chapter 12 introduces the basic principles of the *analysis of variance* (ANOVA). The one-way and two-way analyses of variance are discussed.

The elements of study design are not formally covered in this book but are informally introduced in much of the example material. The concepts of matching, cohort studies, case-control studies, retrospective studies, prospective studies, and the sensitivity, specificity, and predictive value of screening tests are extensively discussed in the context of actual samples. In addition, specific sections on sample size estimation are provided for different statistical situations in Chapters 7, 8, and 10.

A flowchart of appropriate methods of statistical inference on page 575 provides an easy reference to the methods developed in this book. This flowchart is referred to at the end of each of Chapters 6 through 12 to give the student some perspective on how the methods in a particular chapter fit in with the overall collection of statistical methods introduced in this book.

In addition, an index summarizing all examples and problems used in this book is provided, grouped by *medical speciality*.

I am grateful to the Literary Executor of the late Sir Ronald A. Fisher, F.R.S., to Dr. Frank Yates, F.R.S., and to the Longman Group Ltd., London, for permission to reprint Table III from their book *Statistical Tables for Biological, Agricultural and Medical Research* (sixth edition, 1974).

I am indebted to Marie Sheehan and Harry Taplin, who have been invaluable in helping to type this manuscript. Michael Payne, Susan London, and Ian List were also instrumental in providing editorial advice in the preparation of the manuscript. I am grateful to Beow Yeap and Edward Freedman for their assistance in proofreading the manuscript. Finally, I am indebted to my many colleagues at the Channing Laboratory, most notably Edward Kass, Frank Speizer, Charles Hennekens, Frank Polk, Ira Tager, Jerome Klein, James Taylor, Stephen Zinner, Walter Willett, and Alvaro Munoz and to my other colleagues at the Harvard Medical School, most notably Frederick Mosteller, Eliot Berson, Robert Ackerman, Mark Abelson, Leo Chylack, Eugene Braunwald, and Arther Dempster, who provided the inspiration for writing this book.

Bernard Rosner
Boston, MA

Contents

1 General Overview

Statistics is the science whereby inferences are made about specific random phenomena on the basis of relatively limited sample material. The field of statistics can be subdivided into two main areas: mathematical and applied statistics. **Mathematical statistics** concerns the development of new methods of statistical inference and requires a detailed knowledge of abstract mathematics for its implementation. **Applied statistics** concerns the application of the methods of mathematical statistics to specific subject areas, such as economics, psychology, and public health. **Biostatistics** is the branch of applied statistics that concerns the application of statistical methods to medical and biological problems.

A good way to learn about biostatistics and its role in the research process is to follow the flow of a research study from its inception at the planning stage to its completion, which usually occurs when a manuscript reporting the results of the study is published. I will now describe to you one such study in which I participated.

A friend called one morning and in the course of conversation mentioned to me that he had recently used a new, automated blood pressure device of the type seen in many banks, hotels, and department stores. The machine had read his average diastolic blood pressure on several occasions as 115 mm; the highest reading was 130 mm. I was horrified to hear of his experience, since if these readings were true, then my friend might be in imminent danger of having a stroke or developing some other serious cardiovascular disease. I referred him to a clinical colleague of mine, who used a standard blood pressure cuff and measured my friend's diastolic blood pressure as 90 mm. The contrast in the readings aroused my interest, and I began to routinely jot down the readings on the digital display every time I passed the machine at my local bank. I got the distinct impression that a large percentage of the reported readings were in the hypertensive range. Although one would expect that hypertensives would be more likely to use such a machine, I still believed that blood pressure readings obtained with the machine might not be comparable with standard methods of blood pressure measurement. I spoke to Dr. B. Frank Polk about my suspicion and succeeded in interesting him in a small-scale evaluation of such machines. We decided to send a human observer who was well trained in blood pressure measurement techniques to several of these machines. He would offer to pay the subjects 50¢ for the cost of using the machine if they would agree to fill out a short questionnaire and have

their blood pressure measured by both the human observer and the machine.

At this stage we had to make several important decisions, each of which would prove to be vital to the success of the study. The decisions were based on the following questions:

1. How many machines should we test?
2. How many people should we test at each machine?
3. In what order should the measurements be taken—should the human observer or the machine be used first? Ideally, we would have preferred to avoid this problem by taking both the human and machine readings simultaneously, but this procedure was logistically impossible.
4. What other data should we collect on the questionnaire that might influence the comparison between methods?
5. How should the data be recorded to facilitate their computerization at a later date?
6. How should the accuracy of the computerized data be checked?

We resolved these problems as follows:

1. and 2. We decided to test more than one machine (four to be exact), since we were not sure if the machines were comparable in quality. However, we wanted to sample enough subjects from each machine so that we would have an accurate comparison of the standard and automated methods for each machine. We tried to predict how large a discrepancy there might be between the two methods. Using the methods of sample size determination discussed in this book, we calculated that we would need 100 subjects at each site to have an accurate comparison.

3. We then had to decide in what order the measurements should be taken for each person. According to some reports, one problem that occurs with repeated blood pressure measurements is that persons tense up at the initial measurement, yielding higher blood pressures than at subsequent repeated measurements. Thus, we would not always want to use the automated or manual method first, since the effect of the method would get confused with the order-of-measurement effect. A conventional technique that we used here was to **randomize** the order in which the measurements were taken, so that for any person it was equally likely that the machine or the human observer would take the first measurement. This random pattern could be implemented by flipping a coin or, more likely, by using a table of **random numbers** as appears in Table 4 of the appendix.

4. We felt that the major extraneous factor that might influence the results would be body size, since we might have more difficulty getting accurate readings from persons with fatter arms than from those with leaner arms. We also wanted to get some idea of the type of people who use these machines; so we asked questions about age, sex, and previous hypertensive history.

5. To record the data, we developed a coding form that could be filled out on site and from which data could be easily entered on a computer terminal for subsequent analysis. Each person in the study was assigned an identification (ID)

number by which the computer could uniquely identify that person. The data on the coding forms were then keyed and verified. That is, the same form was entered twice, and a comparison was made between the two records to make sure they were the same. If the records were not the same, then the form was reentered.

6. After data entry we ran some editing programs to ensure that the data were accurate. Checking each item on each form was impossible because of the large amount of data. Alternatively, we checked that the values for individual variables were within specified ranges and printed out aberrant values for manual checking. For example, we checked that all blood pressure readings were at least 50 and no more than 300 and printed out all readings that fell outside this range. This simple process enabled us to detect records that had been "offpunched"; that is, at some point in the data entry a column had been skipped, rendering all subsequent data on such records meaningless.

After completion of the data collection, data entry, and data editing phases, we were ready to look at the results of the study. The first step in this process is to get a general feel for the data by summarizing the information in the form of several descriptive statistics. This descriptive material can be numerical or graphical. If numerical, it can be in the form of a few summary statistics, which can be presented in tabular form or, alternatively, in the form of a **frequency distribution**, which lists each value in the data and how frequently it occurs. If graphical, the data are summarized pictorially and can be presented in one or more figures. The appropriate type of descriptive material will vary with the type of distribution considered. If the distribution is **continuous**, that is, if there are essentially an infinite number of possible values, as would be the case for blood pressure, then means and standard deviations might be the appropriate descriptive statistics. However, if the distribution is **discrete**, that is, if there are only a few possible values, as would be the case for sex, then percentages of people taking on each value would be the appropriate descriptive measure. In some cases both types of descriptive statistics are used for continuous distributions by condensing the range of possible values into a few groups and giving the percentage of people that fall into each group (e.g., the percentages of people that have blood pressures between 120 and 129 mm and between 130 and 139 mm).

In this study we decided first to look at mean blood pressure for each method at each of the four sites. Table 1.1 summarizes this information [1].

You might notice from this table that we did not obtain meaningful data from all of the 100 people interviewed at each site, since we could not obtain valid readings from the machine for many of the people. This type of missing data problem is very common in biostatistics and should be anticipated at the planning stage when deciding on sample sizes (which was not done in this study).

Our next step in the study was to determine whether the apparent differences in blood pressure between machine and human measurements at two of the locations (C, D) were "real" in some sense or were "due to chance." This type of question falls into the area of **inferential statistics**. We realized that although there was a 14-mm difference in mean systolic blood pressure between the two methods for the 98 people we interviewed at location C, this difference might not hold up if we interviewed 98 other people at a different time, and we wanted to have some idea

Table 1.1
Mean blood pressures and differences between machine and human readings at four locations.

		Systolic blood pressures (mm Hg)					
		Machine		Human		Difference	
Location	Number of people	Mean	Standard deviation	Mean	Standard deviation	Mean	Standard deviation
A	98	142.5	21.0	142.0	18.1	0.5	11.2
B	84	134.1	22.5	133.6	23.2	0.5	12.1
C	98	147.9	20.3	133.9	18.3	14.0	11.7
D	62	135.4	16.7	128.5	19.0	6.9	13.6

(By permission of the American Heart Association, Inc.)

as to the **error in the estimate** of 14 mm. In technical jargon this group of 98 people represents a **sample** from the **population** of all people who use that machine. We are interested in the population and we wish to use the sample to help us learn something about the population. In particular, we wanted to know how different the **estimated** mean difference of 14 mm in our sample was likely to be from the **true** mean difference in the population of all people who might use this machine. More specifically, we wanted to know if it was still possible that there was no underlying difference between the two methods and that our results were due to chance. We refer to the 14-mm difference in our group of 98 people as an **estimator** of the true mean difference (d) in the population. The problem of inferring characteristics of a population from a sample is the central concern of statistical inference and is a major topic in this text. To accomplish this aim, we needed to develop a **probability model**, which would tell us how likely it is that we would obtain a 14-mm difference between the two methods in a sample of 98 people if there were no real difference between the two methods over the entire population of users of the machine. If this probability were sufficiently small, then we would begin to believe that a real difference existed between the two methods. In this particular case, using a probability model based on the t distribution, we were able to conclude that this probability was less than 1 in 1000 for each of machines C and D. This probability was sufficiently small for us to believe that there was a real difference between the automatic and manual methods of taking blood pressure for two of the four machines tested.

We used a statistical package to perform the preceding data analyses. A package is a collection of statistical programs that describe data and perform various statistical tests on the data. Currently the most widely used statistical packages include SAS, SPSS, BMDP, and MINITAB.

The final step in this study, after completing the data analysis, was to compile the results in the form of a publishable manuscript. Inevitably, because of space considerations, much of the material developed during the data analysis phase was weeded out and only the essential items were presented for publication.

I hope that the review of this study gives you some idea of what medical research is about and what the role of biostatistics is in this process. The material in this text will parallel the description of the data analysis phase of the study